DOCUMENT RESUME

ED 083 281 TM 003 243

AUTHOR	Angoff, William H.; Modu, Christopher C.
TITLE	Equating the Scales of the Prueba de Aptitud Academica and the Scholastic Aptitude Test.
INSTITUTION	College Entrance Examination Board, New York, N.Y.
REPORT NO	CEEB-RR-3
PUB DATE	73
NOTE	44p.
AVAILABLE FROM	Publications Order Office, College Entrance Examination Board, Box 592, Princeton, N.J. 08540 ($1.50)
EDRS PRICE	MF-$0.65 HC Not Available from EDRS.
DESCRIPTORS	*Aptitude Tests; *Comparative Statistics; Cultural Differences; English; *Equated Scores; Language Role; Mathematical Models; *Mathematics; Spanish; Statistical Analysis; Student Testing; Tables (Data); Technical Reports; Test Results; *Verbal Tests
IDENTIFIERS	PAA; Prueba de Aptitud Academica; SAT; Scholastic Aptitude Test

ABSTRACT

The purpose of this study was to establish score equivalencies between the College Board Scholastic Aptitude Test (SAT) and its Spanish-language equivalent, the College Board Prueba de Aptitud Academica (PAA). The method of the study involved two phases: the selection of test items equally appropriate for Spanish- and English-speaking students for use in equating the two tests; and the equating analysis itself. The method of the first phase was to choose two sets of items, one originally appearing in Spanish, the other originally appearing in English; to translate each set into the other language; and to administer both sets in the appropriate language mode for pretest purposes to both types of students. These administrations were conducted in the Fall of 1970 with samples of condidates taking the PAA or the SAT at regularly scheduled administrations. They provided data regarding the difficulty and discrimination power of each item for each of the two groups, and an index of appropriateness of each item for both groups. On the basis of the analyses of these data, two sets of items, one verbal and the other mathematical, were chosen and assembled as "common items" to be used for equating. In the second phase of the study, these "common items," appearing in Spanish and in English, were administered in the appropriate language along with the operational form of the PAA in November 1971 and with the operational form of the SAT in January 1972. The data resulting from the administrations of these "common items" were used to calibrate for differences in the abilities of the two groups and permitted both linear and equipercentile equating of the two tests. Conversion tables are provided. (Author/DB)

FILMED FROM BEST AVAILABLE COPY

ED 083281

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL BY MICROFICHE ONLY HAS BEEN GRANTED BY
Susan Gardner
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER

College Entrance Examination Board Research Report 3

EQUATING THE SCALES OF THE PRUEBA DE APTITUD ACADÉMICA AND THE SCHOLASTIC APTITUDE TEST

William H. Angoff
Christopher C. Modu

The College Entrance Examination Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,000 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the Board's programs and participate in the determination of its policies and activities.

This booklet is one of the College Board Research Reports. This series contains reports of research sponsored by the College Entrance Examination Board that is judged to be of interest to members of the Board, and to other members of the educational community who are not research specialists. The booklets are regular technical reports but are chosen for their importance to those guidance personnel, admissions officers, financial aid directors, teachers, and school and college administrators who may not regularly see professional journals of psychological measurement. Other research supported by the College Board is reported in books and journal articles, or in bulletins and memorandums addressed to specialized audiences. Information about these reports will be supplied upon request.

EC 083281

EQUATING THE SCALES OF THE PRUEBA DE APTITUD ACADÉMICA AND THE SCHOLASTIC APTITUDE TEST

Research Report **3**

William H. Angoff
Christopher C. Modu

EDUCATIONAL TESTING SERVICE

TM 003 243

COLLEGE ENTRANCE EXAMINATION BOARD · NEW YORK
1973

▪ RESEARCH REPORTS

RESEARCH REPORT 1 *Effects of Special Instruction for Three Kinds of Mathematics Aptitude Items*
by Lewis W. Pike and Franklin R. Evans (1972) $1.50.

RESEARCH REPORT 2 *Black Students in Predominantly White North Carolina Colleges and Universities*
by Junius A. Davis and Anne Borders-Patterson (1973) $1.50.

RESEARCH REPORT 3 *Equating the Scales of the Prueba de Aptitud Académica and the Scholastic Aptitude Test*
by William H. Angoff and Christopher C. Modu (1973) $1.50.

Copies of these reports may be ordered from
Publications Order Office,
College Entrance Examination Board, Box 592,
Princeton, New Jersey 08540.

Editorial inquiries concerning these reports
should be directed to *Editorial Office,*
College Entrance Examination Board,
888 Seventh Avenue, New York, New York 10019.

Copyright © 1973 by College Entrance Examination Board.

All rights reserved.

Printed in the United States of America.

This booklet designed by Sophie Adler.

Contents

PREFACE v

ABSTRACT 1

INTRODUCTION 3

METHOD 5

Phase I – Selection of Items for Equating 5

Phase II – Equating 15

Linear Equating 19

Curvilinear Equating 25

SUMMARY AND DISCUSSION 28

REFERENCES 31

APPENDIX

Table I *Linear Conversions of PAA Scaled Scores to SAT Scaled Scores* 35

Table II *Curvilinear Conversions of PAA Scaled Scores to SAT Scaled Scores* 36

Table III *Final Conversions Between PAA Scaled Scores and SAT Scaled Scores* 37

Preface

When in the 1960s the College Board created a test in the Spanish language modeled on the Scholastic Aptitude Test (SAT) the intention was to offer the new test as a form of technical assistance to educators in Puerto Rico and other Spanish-speaking areas who might wish to use this kind of testing to facilitate the transition from school to college. But, as so often happens, the successful development of the Spanish-language Prueba de Aptitud Académica (PAA) has created a variety of uses or proposed uses of the test which were not provided for in the initial planning.

Administrators in continental United States institutions have quite reasonably asked if they can use the PAA as part of the evidence on which to base decisions concerning the admission of Spanish-speaking students. In addition, many educators who are concerned with improving access to higher education for those United States residents and citizens with Spanish backgrounds (Mexican Americans and Puerto Ricans are the largest groups, but there are others) have wondered if aptitude testing in Spanish might not be more appropriate for some of these students than the present English-language testing is.

These are complicated questions which will require time and extensive experimentation to answer. But one of the preliminary steps necessary for beginning to deal with them is to develop methods of equating the Spanish-language PAA and the English-language SAT, so that a particular score on one scale will have the same educational and psychological meaning as a definite score on the other scale.

Equating the several English-language forms of the SAT to one another has been a regular practice for many years and is an important advantage of the Board's program. A student may take any form of the test offered throughout the calendar year with confidence that the score reported to a college will not be influenced by the date of the test, the competition on that particular date, or any other factors except his own ability and, to a slight extent, the inevitable errors of measurement. At first glance, it would seem a relatively simple step to extend these usual equating techniques to the PAA–SAT relationship. But at second glance, the job turns out to be impossible.

The meaning of equated test forms when both are in the same language is in itself not simple. If two forms of a test are properly equated, any given scaled score earned by a student on one form represents the same level of ability or achievement as does the same scaled score earned by another student on another form.

Similarly, if two forms are properly equated, the score earned by a student on one form is most likely the score he would have earned on the other form had he taken that other form at exactly the same time and under exactly the same conditions. These ideas—"the same level of ability" and "the score he would have earned . . . had he taken that other form at exactly the same time"—are difficult enough to think about and to convert into operational terms under ordinary circumstances. But when two languages and cultures are involved, there is no way at all to deal with the idea of level of difficulty with the kind of precision that is ordinarily expected of equating procedures.

But, as occasionally happens, we can admit the impossibility of a task and still undertake to do it as well as we can. The authors in this case have devised an extremely ingenious method for equating the PAA and SAT scales in spite of the difficulties. What they have done has been rather than attempt to force the two tests to yield equated scores, to develop instead tables that show what a particular score on one of the examinations would be equivalent to on the other examination.

It is important to emphasize, as the authors do in the text, that the equating tables thus provided were developed for one particular population—Puerto Rican students in Puerto Rico—and are not known to be accurate for, say, Cubans or Colombians or Puerto Ricans in New York. Certainly, the suitability of the PAA (plus equating tables) for a Chicano student in Los Angeles is a question even further removed from the data of this research.

It is also important to say that equating the PAA to the SAT does not confer validity upon the PAA in circumstances where the SAT is known to give useful predictions of college performance. Each test should be validated by studying its usefulness in each institution where its use is contemplated.

This equating experiment must not be thought of only from the point of view of the SAT and the PAA. It is an important step forward in examining the problem of cross-cultural testing, and may have useful applications in many other settings where students must either study in languages different from the ones spoken in their homes, or where there is some other reason to question the equivalence of scores across linguistic and cultural distances.

The educator who must use test scores may use the tables provided here to do his work more effectively—but always with the caution indicated by the limitations of the study. In addition, the student of testing and the connoisseur of equating will find this a splendid example of his science and art.

S. A. Kendrick

Chief, Division of Research Studies and Services
College Entrance Examination Board

Abstract

The purpose of this study was to establish score equivalencies between the College Board Scholastic Aptitude Test (SAT) and its Spanish-language equivalent, the College Board Prueba de Aptitud Académica (PAA). The method of the study involved two phases: the selection of test items equally appropriate for Spanish- and English-speaking students for use in equating the two tests; and the equating analysis itself. The method of the first phase was to choose two sets of items, one originally appearing in Spanish, the other originally appearing in English; to translate each set into the other language; and to administer both sets in the appropriate language mode for pretest purposes to both types of students. These administrations were conducted in the fall of 1970 with samples of candidates taking the PAA or the SAT at regularly scheduled administrations. They provided data regarding the difficulty and discrimination power of each item for each of the two groups, and, what was of special interest, an index of appropriateness of each item for both groups.

On the basis of the analyses of these data, two sets of items, one verbal and the other mathematical, were chosen and assembled as "common items" to be used for equating. In the second phase of the study these "common items," appearing in Spanish and also in English, were administered in the appropriate language along with the operational form of the PAA in November 1971 and with the operational form of the SAT in January 1972. The data resulting from the administrations of these "common items" were used to calibrate for differences in the abilities of the two groups of candidates and permitted both linear and equipercentile equating of the two tests. Conversion tables relating the PAA-verbal scores to the SAT-verbal scores and the PAA-mathematical scores to the SAT-mathematical scores are given in the Appendix (pages 35–37). These conversions represent an average of the linear and equipercentile results. Because of the scarcity of data at the upper end of the distribution of PAA scores, score equivalencies are permissible, strictly speaking, only as high as the mid-700s. Score equivalencies beyond the mid-700s were obtained by extrapolation.

Introduction

Although the study of cultural differences has been of central interest to educational and social psychologists for a long while, attempts to develop a deeper understanding of this area have been frustrated by the absence of a common metric by which such comparisons could be made. The reasons for this are obvious. If two groups differ from each other in ways that cast doubt on the validity of any direct comparisons between them—if, for example, they differ in language, customs, and values—then any problem that defies direct comparisons also defies the construction of an unbiased metric by which one could hope to make those comparisons.

The present study represents an attempt to develop a methodology to help make comparisons in the face of these difficulties, and to provide a conversion of the verbal and mathematical scores on the Spanish-language Prueba de Aptitud Académica (PAA) of the College Board to the verbal and mathematical scores, respectively, on the College Board English-language Scholastic Aptitude Test (SAT). Both tests, it is to be noted, are administered to secondary school students for admission to college. The PAA is typically administered to Puerto Rican students who are planning to attend colleges and universities in Puerto Rico; the SAT is typically administered to mainland students who are planning to attend colleges and universities in the continental United States. It was expected that if conversion tables between these two tests (and score scales) were made available, direct comparisons could be made between subgroups of individuals of the two language-cultures who had taken only that test appropriate for them. It was also expected that these conversion tables would help in the evaluation of the probable success of Puerto Rican students who were interested in eventually attending colleges on the mainland and were submitting PAA scores for admission.

Interest in developing conversions such as these has been expressed in various other contexts, usually in the assessment of the outcomes of education for differ-

This project was supported by the College Entrance Examination Board. The authors wish to express their deep appreciation to Dr. E. Elizabeth Stewart for her many helpful comments and suggestions in the review of this manuscript; to the staff of the Foreign Language Department of Educational Testing Service, who coordinated and participated in the translation of the items for this study; and to Mr. Carlos J. Lopez-Nazario and the staff of the College Board Puerto Rico Office for their encouragement, cooperation, and able assistance throughout the entire course of the study.

ent cultural groups living in close proximity—for English- and French-speaking students in Canada, for example; for English- and Afrikaans-speaking students in South Africa; for speakers of one or another of the many languages in India or in Africa, etc. However, no satisfactory methods to satisfy this interest have been evident, and the problems attendant on making comparisons among culturally different groups are far more obvious and numerous than are the solutions. For example, in order to provide a measuring instrument to make these comparisons, it is clearly insufficient simply to translate the test constructed for one language-group into the language of the other, even with adjustments in the items to conform to the more obvious cultural requirements of the second group. It can hardly be expected, without making careful and detailed checks—assuming that such checks can logically be made—that the translated items will have the same meaning and relative difficulty for the second group as they had for the original group before translation.

A method considerably superior to that of simple translation has been described by Boldt (1969). It requires the selection of a group of individuals who are judged to be equally bilingual and bicultural, and the administration of two tests to each individual, one in each of the two languages. Scores on the two tests are then equated as though they were parallel forms of the same test, and a conversion table is developed relating scores on each test to scores on the other.

One of the principal difficulties with the foregoing procedure is that the judgment "equally bilingual and bicultural" is an extremely difficult, perhaps even an impossible, one to make. More than likely, the group is more proficient, on the average, in one of the two languages than in the other. This would be especially true, of course, if the group is constituted from a small number of clusters of individuals.

The present study represents an attempt to overcome such difficulties. In brief, it calls for administering the PAA to Puerto Rican students and the SAT to mainland U.S. (continental) students, using a set of "common," or anchor, items to calibrate and adjust for any differences between the groups in the process of equating the two tests. It is noted that these items are "common" only in terms of the operations used to develop and select them. By the very nature of things they had to be administered in Spanish to the Puerto Rican students and in English to the continental students. Therefore, to the extent that there is any validity in the notion that a set of test items can represent the same psychological task to individuals of two different languages and cultures, to the extent that the sense of the operations is acceptable, and to the extent that the operations themselves were adequate, the study will have achieved its purpose. There is also the concern that the Puerto Rican and continental groups appear to differ so greatly in average ability that with the limited equating techniques available it is not likely that any set of common items, however appropriate, can make adequate adjustments for the differences, even if the two tests were designed for students of the same language and culture.

There is, finally, the concern about the generalizability of a conversion between tests that are appropriate for different cultural groups. In the usual equating prob-

lem, a conversion function is sought that will simply translate scores on one form of the test to the score scale on a parallel form of the test – an operation analogous to that of translating Fahrenheit units of temperature to centigrade units. However, when the two tests in question are measuring different types of abilities, or when one or both of the tests may be unequally appropriate for different subgroups of the population, the conversion cannot be unitary, as would be true of the temperature-scale conversion, but would be different for different subgroups (Angoff, 1966). In the case of the present equating attempt, it is entirely possible that the use of different types of subgroups for the equating experiment – Mexicans and Australians, for example, instead of Puerto Ricans and U.S. continentals – would yield conversion functions quite different from those developed in the present study. For this reason the conversions developed here should be considered as having limited applicability, and should not be used without verification with groups of individuals much different from those studied here.

Method

The method followed in this study for deriving conversions of scores from the verbal and mathematical scales of the PAA to the verbal and mathematical scales of the SAT consisted of two phases. The first phase entailed the selection of appropriate anchor items for the equating analysis. This phase involved the preparation of sets of items in Spanish and in English; the translation of each set into the other language; and the administration of both sets in the appropriate language to both Spanish- and English-speaking students. On the basis of an item analysis of the data resulting from this administration, groups of verbal and mathematical items were chosen to fulfill the principal requirement that they be equally appropriate, insofar as this could be determined, for both groups of students. Beyond this, the usual criteria for the choice of equating items as to difficulty, discrimination, and content coverage were adhered to wherever possible. Once the anchor items were chosen, the second phase was undertaken, calling for a second test administration and an analysis for equating, based on the data resulting from that administration.

Phase I – Selection of Items for Equating

In accordance with this plan, 58 Spanish verbal items, 97 English verbal items, 48 Spanish mathematical items, and 52 English mathematical items were chosen from the files. (An effort was made to assemble equal numbers of items in Spanish and English, but the pool of pretested and usable Spanish items, particularly verbal items, did not permit this.) Each item was translated by a small team of bilingual experts into the other language, thus making available two complete sets of items, 155 verbal and 100 mathematical, each set appearing in both languages and, as nearly as was possible by translation, equally meaningful in both languages.

At a later time, all the items ultimately selected as anchor items for equating were retranslated independently by different translators back into the original

languages. When the original version of each item was compared with the version that had undergone two translations—from the original language (Spanish or English) to the other language (English or Spanish), and back again to the original language—it was found that the two generally compared very well, indicating that the translation was adequate and that the original meaning of most of the items seemed to have undergone no great change through the course of these two translations, a consideration fundamental to the success of this study.

The 155 verbal items consisted of four types: antonyms, analogies, sentence completion, and reading comprehension. The 100 mathematical items were of two types: arithmetic and algebraic reasoning problems and problems involving geometric concepts. Detailed information on the pretested items is given later in this report.

The 155 verbal items and the 100 mathematical items were each subdivided into subsets of items and administered to systematic samples of regular College Board examinees. The items appearing in Spanish were taken by candidates for the Spanish-language PAA at the November 1970 administration of the PAA; the same items, appearing in English, were taken by candidates for the English-language SAT at the November 1970 administration of the SAT. Five systematic samples of Puerto Rican candidates (4 of 305 cases and 1 of 310 cases) were formed, each taking 1 of 5 subsets of 31 verbal items in a 25-minute testing period. Five additional Puerto Rican samples (4 of 270 cases and 1 of 275 cases) were similarly formed, each taking 1 of 5 subsets of 20 mathematical items in a 25-minute testing period. Correspondingly, 8 systematic 2,000-case samples of continental (United States) candidates were formed, each taking 1 of 4 subsets of 40 verbal items[1] or 1 of 4 subsets of 25 mathematical items in a 30-minute testing period.

Since the five sets of Spanish verbal items and the five sets of Spanish mathematical items were administered to different, although very similarly performing, groups of Puerto Rican students, minor equating adjustments were made in the difficulty indexes so that comparisons across the sets of items within each domain (verbal and mathematical) could be made directly. The method of adjustment was essentially that described by Thurstone (1947). The same types of adjustments were made for the items in the four verbal and four mathematical sets administered in English. Once these adjustments were carried out it was possible to pool all the verbal items appearing in Spanish into one undifferentiated set and all the verbal items appearing in English into a second undifferentiated set and prepare for the next step in the analysis. (All the mathematical items in each language were similarly pooled into one total set.) This step, which consisted of an examination and comparison of the performance of the Puerto Rican students with the performance of the continental students on the "same" items, involved a procedure which requires detailed description.

In preparation for making this comparison, the proportion p in each of the two

[1] In order to permit the formation of four subsets of 40 items each, five "filler" verbal items were added to the 155, making a total of 160 items.

language-groups answering each item correctly is calculated and converted to Δ.[2] A plot is then made of the points represented by the paired Δ-values, Δ_{g_i} vs. Δ_{h_i}, where g represents one of the groups and h the other, one point for each of the items i under consideration for which Δ-values are available. The plot of these points is normally an ellipse extending from lower left to upper right, and if the samples are drawn from the same types of populations, the scatterplot of these points is a long, narrow one, often representing a correlation as high as .98 or .99. When the samples are somewhat different in level, the points still fall in a long narrow ellipse, but it is displaced vertically or horizontally, depending on which group is the abler one. Even when the groups differ in dispersion the points still fall in the same type of ellipse, but it is tilted at an angle either smaller or larger than 45°, depending on which sample is more dispersed. However, when the groups differ in type, or when the items do not all have the same meaning for the two groups—which may often be the case when the groups are drawn from the same general type of population but differ sharply in level or dispersion—the item difficulties will not fall in precisely the same rank order for the two groups, and the correlation represented by the delta points will be lower than .98 or .99, sometimes substantially lower. The items falling at some distance from the plot may be regarded as contributing to the item-by-group interaction. They are the items that are especially more difficult for one group than for the other, relative to the other items, and they are the items that appear to represent different "psychological meanings" to the members of the two groups.

The purpose of the delta plots is to enable the identification of those items that do in fact have different meaning for the two groups. The method developed to accomplish this involves the determination of the major axis of the ellipse formed by the plotted points, and the calculation of the perpendicular distance D_i from each point to the line. If there were no other consideration in the choice of items, the items represented by the smallest D_i-values would be retained; the others would be eliminated.

The equation used for the major axis of the ellipse is a linear one, $h = Pg + Q$, where

$$P = \frac{(s_h^2 - s_g^2) \pm \sqrt{(s_h^2 - s_g^2)^2 + 4r_{gh}^2 s_g^2 s_h^2}}{2r_{gh}s_g s_h}$$

and

$$Q = M_h - PM_g.$$

(The variables g and h are, respectively, the delta values for the two groups under consideration.) The formula for the perpendicular distance D_i of each point i in the plot to the line is given as:

$$D_i = \frac{Pg_i - h_i + Q}{\sqrt{P^2 + 1}}.$$

[2] $\Delta = 4z + 13$, where z is a normal deviate corresponding to p; Δ is inversely related to p, the higher the delta-value the more difficult the item.

Items were defined as "equally appropriate" to the Spanish- and English-speaking groups on the basis of their proximity to the major axis of the delta plot. If the ellipse itself is biased toward one group or the other, then the items chosen as "equally appropriate" to both groups will also tend to be biased toward that group. Thus, in addition to the fact that the item-by-group interaction between Spanish and English speakers is far greater, as will be observed below, than would be ideal for equating the scales of the tests appropriate for these groups, it should be noted that the final conversion may still contain some elements of bias in spite of the fact that the method of choosing "equally appropriate" items is intended to eliminate the major sources of bias.

Following the administrations of the items in their Spanish and English forms, indexes of item difficulty (deltas) and discrimination (biserial correlations with the operational verbal or mathematical score, as appropriate) were calculated separately for the two language-groups. A plot was then made of the delta-value for each item observed in the PAA group vs. the delta-value for that same item observed in the SAT group and a measure of the item-by-group interaction D_i for that item (defined in the preceding section as the perpendicular distance of each point representing the paired delta-values for each item i from the major axis of the bivariate ellipse) was also calculated. These three indexes formed the principal basis for the final selection of the 40 verbal and 25 mathematical items to be used as "quasi-common" ("anchor") items for the equating of the two scales. Items which were closest to the major axis of the ellipse and which were also within the limits set for the difficulty and discrimination indexes were the ones used for this purpose.

Figure 1 gives the delta plot for the 155 verbal items, and Figure 2 gives the delta plot for the 100 mathematical items. Points to the right of the major axis in each of these ellipses represent items that were more difficult, relative to the other items, for the SAT group than for the PAA group. These items are represented by positive D-values. Points to the left of the major axis represent items that are more difficult, relative to the other items, for the PAA group, and are represented by negative D-values. Note (see Figure 1) that 21 of the 155 verbal items were harder for the SAT group than for the PAA group. Fifteen of these items were originally in English; the other six were originally in Spanish. The remaining 134 verbal items were more difficult for the PAA group. As may be seen in Figure 2, all 100 mathematical items—the 52 originally in English and the 48 originally in Spanish—were considerably more difficult for the PAA group.

The plot of verbal items in Figure 1 is far more dispersed about the major axis than is the corresponding plot of mathematical items in Figure 2, indicating a much lower correlation for verbal items (.60) than for mathematical items (.85). As was pointed out earlier in this report, the correlation between Δ-values may be regarded as a measure of item-by-group interaction. In those instances where the two groups are drawn from the same general population it is not unusual to find correlations in the neighborhood of .98 and even higher. The fact that these correlations, particularly the correlation for the verbal items, are as low as they are suggests that the items do not have quite the same psychological meaning for

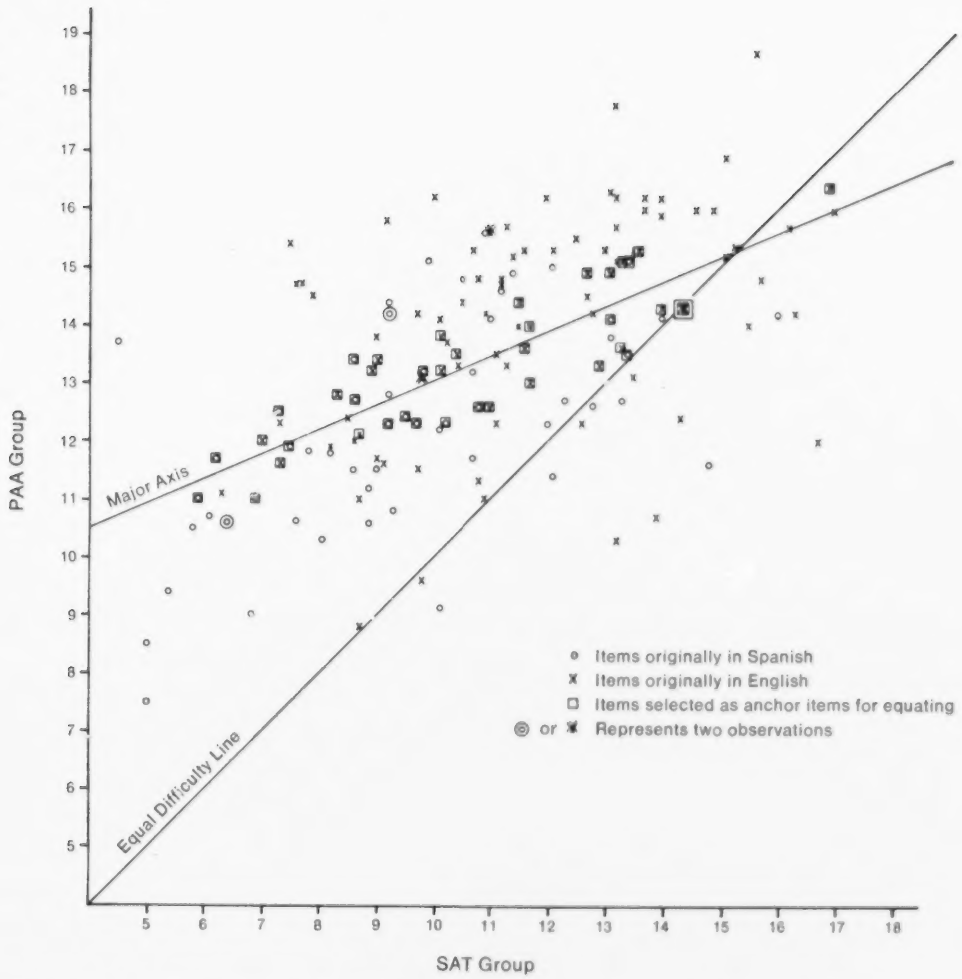

FIGURE 1. *Delta Plot for the Pretested Verbal Items (Number of Items = 155)*

the members of these two language-groups. In a sense this is one of the most significant findings in the present study, since it reflects in the form of statistical data the very nature of the psychological difficulties that are likely to be encountered in making cross-cultural studies. With respect to this study in particular, it casts some doubt on the quality of any equating that would be carried out with these items. Since the equating items are used to calibrate for differences in the abilities of the PAA and SAT groups, a basic requirement for equating is that they have the same rank order of difficulty in the two groups. Considerable improvement, in the sense of reducing the item-by-group interaction, was achieved in the group of verbal items, as will be shown below, by discarding the most aberrant ones among them. Nevertheless, with item-by-group interaction effects as large as those observed here, the concern remains that the equating might be much less trustworthy than would be expected of an equating of two parallel tests intended for members of the same language-culture.

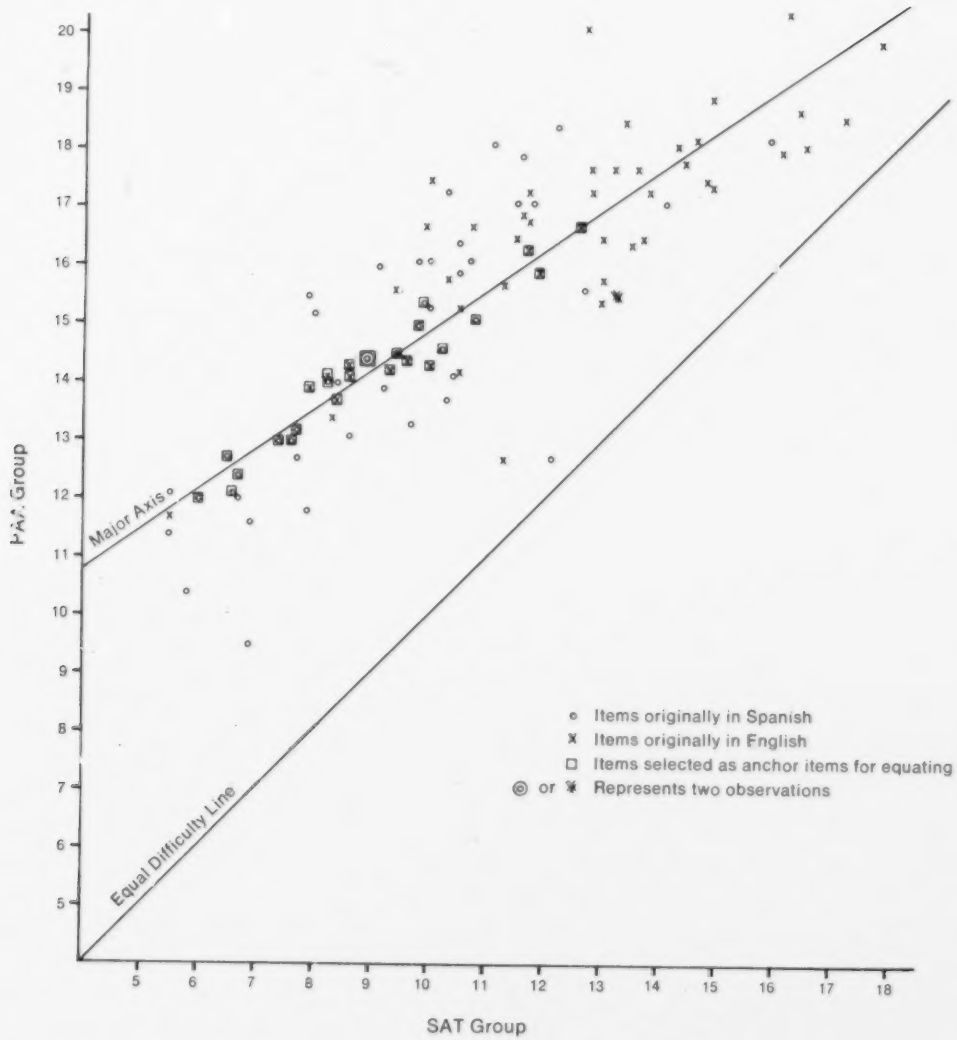

FIGURE 2. *Delta Plot for the Pretested Mathematical Items (Number of Items = 100)*

It should be reiterated, however, that these interactions were not entirely unexpected; the observation has often been made that verbal material, however well it may be translated into another language, loses some of its subtleties in the process of translation. Even in the case of mathematical items some shift in the order of item difficulty is to be expected, possibly because of differences between Puerto Rico and the United States mainland with respect to the organization and emphasis of the mathematics curriculum in the early grades.

Summary statistics—means and standard deviations (SD)—indexes (Δ, r_{bis}, and D) that were used as a basis for identification and selection of the equating items are given in Table 1. These statistics are presented for all the items pretested in this study and also for those finally selected for equating. From the data given in

TABLE 1 *Summary Statistics for the Pretest Items, by Language of Origin, Before and After Selection of Equating Items*

All Pretested Items	Number of Items	Delta-Values (Δ)				Biserial Correlations (r_{bis})				Distances (D) from Major Axis		Correlations Between Deltas
		Mean		SD		Mean		SD				
		SAT	PAA	SAT	PAA	SAT	PAA	SAT	PAA	Mean	SD	
Verbal												
Originally Spanish	58	9.55	12.31	2.64	1.79	.37	.40	.16	.15	.52	1.33	.60
Originally English	97	11.64	14.03	2.53	1.81	.44	.30	.12	.18	−.31	1.48	.48
All Verbal Items	155	10.86	13.38	2.76	1.98	.41	.33	.14	.17	.00	1.48	.60
Mathematical												
Originally Spanish	48	9.25	14.41	2.23	2.13	.53	.45	.09	.17	−.01	1.04	.81
Originally English	52	12.20	16.43	2.65	1.89	.53	.27	.10	.25	.02	.91	.83
All Mathematical Items	100	10.78	15.46	2.86	2.25	.53	.35	.10	.23	.00	.98	.85
Items Selected for Equating[a]												
Verbal	40	10.65	13.29	2.61	1.21	.46	.37	.10	.10	.02	.55	.87
Mathematical	25	8.90	14.13	1.69	1.19	.56	.54	.07	.13	.01	.27	.97

[a] All the statistics in this table are based on the item analysis samples. Independent samples of examinees tested a year later, drawn for the equating phase of this study, showed a correlation between deltas of .81 for the selected verbal items and a correlation of .80 for the selected mathematical items.

Table 1 it is again clear that these items are far more difficult (and less variable in difficulty) for the PAA group than for the SAT group, and that the difference in average difficulty is even more pronounced for the mathematical items than for the verbal items. There also appears to be some tendency for the verbal items to become slightly more difficult after translation. Evidence of this may be found in the column of mean *D*-values, which shows that the verbal items originally in Spanish and translated into English are harder, relative to the other items, for the SAT group. The opposite is true for the items originally in English and translated into Spanish. These were harder, relative to the other verbal items, for the PAA group. No such observation is made in the group of mathematical items.

Generally speaking, the items, especially the mathematical items, have lower discrimination values (r_{bis}) in Spanish than in English. It is quite possible that, as a result of the greater absolute difficulty of the items for the PAA group, those students guessed more frequently than did the continental U.S. students, thereby introducing more error into the items and depressing the item biserials. It is also possible that the items lost in discriminating power as a result of translation. The mean biserial for the verbal items appearing originally in Spanish dropped slightly when translated into English (.40 to .37). Those that appeared originally in English dropped sharply when translated into Spanish (.44 to .30). In the mathematical sections the items originally in Spanish gained when translated into English (.45 to .53); but those that originally appeared in English lost considerably when translated into Spanish (.53 to .27).

The verbal items selected for equating had, for the PAA group, a mean biserial considerably lower than the mean biserial found for the operational PAA-verbal form with which these items were later to be administered for equating (.37 as compared with .45) and, typically, for other operational forms of the PAA. The mean biserials of the selected equating items for the SAT-verbal, SAT-mathematical, and PAA-mathematical items, however, are well in line with the values for the operational forms (.46 as compared with .47; .56 as compared with .54; and .54 as compared with .53; respectively).

Perhaps most interesting in this cross-cultural study are the sizes of the correlations represented by the two delta plots. In the entire group of 155 verbal items the correlation between deltas was only .60, suggesting that as a group these items did not represent the same meaning to the two groups of students. It is noted, however, that the items chosen for equating were much superior in this regard. They represent a correlation of .87 for the same set of data. This comes as no surprise, of course, since only those items with *D*-values closest to the major axis of the plot, within the limits of ± 0.8, were chosen for equating. When this correlation was recalculated in a pair of independent samples (the ones later chosen for use in equating the verbal tests), it dropped, as expected, to .81.

As expected, the correlation between the deltas for the 100 mathematical items was much higher than for the 155 verbal items: .85 as against .60. After selection, the correlation rose to .97, based on the same set of data. This rise in the correlation between deltas was also expected in the mathematical test, since the items selected for equating were those that had *D*-values within the limits ± 0.5. How-

TABLE 2 *Summary Statistics for the Pretest Items, by Item Type*

All Pretested Items	Number of Items	Delta-Values (Δ)				Biserial Correlations (r_{bis})				Distances (D) from Major Axis		Correlations Between Deltas
		Mean		SD		Mean		SD				
		SAT	PAA	SAT	PAA	SAT	PAA	SAT	PAA	Mean	SD	
Verbal												
Antonyms	39	11.50	13.23	3.20	1.86	.45	.39	.13	.16	.37	1.71	.35
Analogies	36	10.51	13.79	2.11	1.57	.41	.34	.12	.14	−.59	1.27	.45
Sentence Completion	40	9.95	12.76	2.90	2.46	.42	.32	.17	.18	.26	1.56	.81
Reading Comprehension	40	11.47	13.79	2.34	1.70	.38	.29	.12	.20	−.16	1.10	.75
Mathematical												
Regular Mathematics	70	10.60	15.32	2.94	2.28	.53	.35	.08	.24	.01	.98	.86
Geometry	30	11.19	15.77	2.65	2.14	.54	.36	.13	.22	−.02	.97	.83

ever, when this correlation was recalculated in a pair of independent samples (those later chosen for use in equating the mathematical tests), it dropped to .80. Although a drop in correlation of some magnitude was expected here too, it was not anticipated that the drop would be so severe. An examination of this delta plot revealed that four items were markedly (but unexplainably) aberrant, two of them extremely so. The removal of these two items raised the correlation to .86; the removal of all four raised it to .94.

Table 2 is a summary of the same data as shown in Table 1, but classified by item type rather than by language of origin. The greater difficulty of the items for the PAA group is readily observable in this table, as is the smaller dispersion of item deltas, not only over the entire test, but also separately by item type. It is also clear that the items in all four of the verbal and in both of the mathematical item types are more discriminating for the continental students than for the Puerto Rican students.

Although the item groups, and the item analysis samples they were based on, are too small to permit easy generalization, it appears that there is considerable and, very likely, significant group-by-item-type interaction, that is to say, variation from one verbal item type to another with respect to their average departure $\bar{D}_i$ from the "equal appropriateness" line (major axis). (No such interaction is observed in the mathematical items.) The analogy items especially, and to some extent the reading-comprehension items, were more difficult, relative to the other items, for the Puerto Rican students than for the continental students. The antonyms and sentence completion items, on the other hand, were relatively less difficult for the Puerto Rican than for the continental students. This appears to be a

TABLE 3 *Distribution of Pretested Items, by Item Type and Language of Origin*

Verbal

	Originally English	Originally Spanish	Totals
Antonyms	24	15	39
Analogies	20	16	36
Sentence Completion	23	17	40
Reading Comprehension	30	10	40
	97	58	155

$\chi^2 = 3.852$; $df = 3$ $\quad .30 > P > .20$

Mathematical

	Originally English	Originally Spanish	Totals
Regular Mathematics	37	33	70
Geometry	15	15	30
	52	48	100

$\chi^2 = .069$; $df = 1$ $\quad .80 > P > .70$

subtle effect, very likely characteristic of the item type itself. It is certainly not a function of the origin of these items and their increased relative difficulty upon translation into the other language. As Table 3 shows, very nearly the same proportion of items for each of the item types was drawn from each of the languages.

It is interesting that the four verbal item types arrange themselves into two distinct classes insofar as the correlations between their deltas are concerned, the higher correlations (smaller item-by-group interactions) characteristic of the sentence-completion and reading-comprehension plots, and the lower correlations (larger item-by-group interactions) characteristic of the antonyms and analogies plots. This result is intuitively reasonable since items with more context tend to retain their meaning, even in the face of translation into another language.

In spite of their resistance to the effects of translation, none of the reading comprehension items was used for the verbal equating test. The reason for this is that the reading comprehension items are not discrete like the others, but are interrelated in groups of five, each group based on a single reading passage. Although one or another of the items within a group may have passed the requirements for use in the equating test, other items in the group did not; and in each group there were enough unusable items to render the whole group unusable.

Phase II—Equating

Once the 40 verbal and 25 mathematical items that were to be used as "common" —more properly, "quasi-common"—items were chosen, preparations were made to administer them to groups of candidates taking the PAA or the SAT for admission to college. Accordingly, two samples of candidates were chosen from the November 1971 administration of the PAA, one to take the verbal items in Spanish, the other to take the mathematical items in Spanish, in addition to the regular operational form of the PAA given at that time. Similarly, two samples of candidates were chosen from the January 1972 administration of the SAT, one to take the verbal items in English, the other to take the mathematical items in English, in addition to the regular operational form of the SAT given at that time. The PAA samples were chosen to represent groups somewhat higher scoring (mean verbal score = 494; mean mathematical score = 488) than the general 1971–72 PAA candidate group (mean verbal score = 478; mean mathematical score = 484); the SAT samples, drawn systematically from the January 1972 administration, were somewhat lower scoring (mean verbal score = 424; mean mathematical score = 473) than the general 1971–72 SAT candidate group (mean verbal score = 450; mean mathematical score = 482). Arrangements were made to administer the equating items in a separately timed 30-minute period to all candidates (both Puerto Rican and continental) taking those items.

The method of equating involved first the determination of the conversion between the raw $[R - (W/4)]$ scores on the operational form of the PAA given in November 1971 and the raw $[R - (W/4)]$ scores on the operational form of the SAT given in January 1972. The second step called for substituting into that conversion relationship the equation between raw and scaled scores for each of the

two operational forms. The result of this work would be the score-to-score relationship between scaled scores in the two testing programs.

Two types of equating were undertaken, linear equating and curvilinear (equipercentile) equating, and within the general linear model two methods were used. The first of these two linear methods, following a procedure described by Tucker (in Angoff, 1971, p. 580), required an estimation of the mean and variance for the combined PAA–SAT samples on each of the tests, as follows:

$$\hat{M}_{x_t} = M_{x_\alpha} + b_{xv_\alpha}(M_{v_t} - M_{v_\alpha}), \tag{1}$$

$$\hat{M}_{y_t} = M_{y_\beta} + b_{yv_\beta}(M_{v_t} - M_{v_\beta}), \tag{2}$$

$$\hat{s}^2_{x_t} = s^2_{x_\alpha} + b^2_{xv_\alpha}(s^2_{v_t} - s^2_{v_\alpha}), \tag{3}$$

and

$$\hat{s}^2_{y_t} = s^2_{y_\beta} + b^2_{yv_\beta}(s^2_{v_t} - s^2_{v_\beta}), \tag{4}$$

where x or X = the test (PAA) taken by group α (Puerto Rican); y or Y = the test (SAT) taken by group β (continental U.S.); v = the score on the "common items," i.e., scores on the items taken in Spanish by the Puerto Rican candidates and scores on the "same" items taken in English by the continental U.S. candidates; and t = the combined group $\alpha + \beta$.

The notation b_{xv} represents the usual coefficient of regression of variable x on variable v: $b_{xv} = r_{xv}s_x/s_v$. (Similarly, $b_{yv} = r_{yv}s_y/s_v$.) The estimated values $\hat{M}_{x_t}$, $\hat{M}_{y_t}$, $\hat{s}_{x_t}$, and $\hat{s}_{y_t}$ are then substituted in the equation

$$\frac{Y - \hat{M}_{y_t}}{\hat{s}_{y_t}} = \frac{X - \hat{M}_{x_t}}{\hat{s}_{x_t}} \tag{5}$$

to yield the equation

$$Y = aX + b, \tag{6}$$

converting the scale of the raw scores on test X to the scale of the raw scores on test Y. In this equation $a = \hat{s}_{y_t}/\hat{s}_{x_t}$ and $b = \hat{M}_{y_t} - a\hat{M}_{x_t}$.

The second type of linear equating, due to Levine (1955), is based on the conversions of true rather than observed scores, and is applicable when the tests (X and Y) to be equated are unequally reliable. In this procedure, when the equating test V is exclusive and experimentally independent of X and Y, the slope a' and intercept b' of the conversion equation

$$Y = a'X + b' \tag{7}$$

are calculated as follows:

$$a' = n_{yv_\beta}/n_{xv_\alpha} \tag{8}$$

and

$$b' = \hat{M}_{y_t} - a'\hat{M}_{x_t}, \tag{9}$$

where

$$\hat{M}_{x_t} = M_{x_\alpha} + n_{xv_\alpha}(M_{v_t} - M_{v_\alpha}), \tag{10}$$

$$\hat{M}_{y_t} = M_{y_\beta} + n_{yv_\beta}(M_{v_t} - M_{v_\beta}), \tag{11}$$

and where, in general, n_{ij} is the ratio of effective test length of test i to test j (Angoff, 1953):

$$n_{ij} = \frac{s_i^2 + r_{ij}s_i s_j}{s_j^2 + r_{ij}s_i s_j}. \tag{12}$$

The derivation of the conversion from the PAA-verbal reporting scale to the SAT-verbal reporting scale is developed as follows. The linear equation

$$S_p = AX + B \tag{13}$$

is the equation by which raw scores on the form of the PAA used in November 1971 (X) are converted to the PAA reporting scale (S_p). Similarly, the linear equation

$$S_c = A'Y + B' \tag{14}$$

is the equation by which raw scores on the form of the SAT used in January 1972 (Y) are converted to the SAT scale (S_c). Expressing equation (13) in terms of X $[X = (S_p - B)/A]$ and equation (14) in terms of Y $[Y = (S_c - B')/A']$, and substituting in equation (6) results in the equation

$$\frac{S_c - B'}{A'} = a\left[\frac{S_p - B}{A}\right] + b,$$

which, when simplified, becomes

$$S_c = \frac{aA'}{A}S_p + A'b + B' - \frac{aA'B}{A}. \tag{15}$$

Equation (15) is a linear equation with slope equal to aA'/A and intercept equal to $A'b + B' - (aA'B/A)$, and may be used to convert verbal or mathematical scores from the November 1971 converted-score scale for the PAA to corresponding scores on the SAT scale.

The curvilinear, or equipercentile, equating between raw scores on the PAA and the SAT followed a procedure described by Angoff (1971, p. 583) involving the following steps: (1) equating the scores on the operational form of the PAA to the scores on the Spanish version of the "common items"; (2) equating the scores on the operational form of the SAT to the scores on the English version of the "common items"; and (3) setting equivalent scores on the PAA and SAT that were found to be equivalent to the same "common item" scores.

TABLE 4 *Frequency Distributions and Summary Statistics for the Operational and Equating Sections of the SAT and PAA*

	Verbal Tests			
	Continental Sample		Puerto Rican Sample	
Raw (Formula) Score	Operational SAT	Equating Section	Operational PAA	Equating Section
84 – 86	2			
81 – 83	3			
78 – 80	6			
75 – 77	10			
72 – 74	22			
69 – 71	30			
66 – 68	45		1	
63 – 65	53		4	
60 – 62	63		17	
57 – 59	99		24	
54 – 56	106		53	
51 – 53	129		68	
48 – 50	131		72	
45 – 47	180		72	
42 – 44	180		84	
39 – 41	205	13	99	
36 – 38	223	67	72	5
33 – 35	250	177	103	21
30 – 32	248	263	92	34
27 – 29	265	375	106	56
24 – 26	238	427	99	92
21 – 23	206	428	91	93
18 – 20	247	495	94	165
15 – 17	187	423	62	122
12 – 14	163	365	61	149
9 – 11	148	321	61	168
6 – 8	112	195	19	163
3 – 5	110	148	17	184
0 – 2	73	73	10	77
−3 – −1	36	22	3	41
−6 – −4	19	6	1	15
−9 – −7	4			
−12 – −10	3			
Number of Cases	3,798	3,798	1,385	1,385
Mean	31.4334	19.4479	32.1047	13.2448
SD	17.3455	8.7787	14.2739	8.8839
Correlation: Operational vs. Equating	.8833		.8488	
Number of Items	90	40	70	40

NOTE: The operational PAA-verbal and SAT-verbal tests were 70 and 90 items respectively. The verbal equating test, administered to both the PAA and SAT groups (in their own language mode), consisted of 40 items.

Linear Equating

In order to calculate the estimated values (for the verbal tests) for the Tucker equating given in equations (1) through (4) and the values for the Levine equating given in equations (8) through (11), the correlations between the operational test and the 40-item equating section,[3] as well as the related means and standard deviations, were prepared for each of the two verbal samples, one consisting of 3,798 continental students and the other consisting of 1,385 Puerto Rican students. These statistics, accompanying the frequency distributions of the operational and equating verbal tests, are given in Table 4.

The data of Table 4 make it clear that, to the extent that the "common items" are in fact appropriate for both groups of examinees, the continental sample is the higher-scoring of the two, by about 0.7 standard deviations. Additional observations may be made regarding the operational tests: The 70-item PAA appears to be only slightly too difficult, on the average, for the Puerto Rican sample; the average percentage-pass on that test (corrected for guessing) was .46. The 90-item SAT is clearly difficult for the continental sample; the average percentage-pass on that test (also corrected for guessing) was .35. These observations are confirmed by the shapes of the distributions in Table 4, which, except for the distribution of the equating-section scores for the continental sample, are all positively skewed.

The patterns of standard deviations and correlations observed in Table 4 between the equating test in English and the SAT and between the equating test in Spanish and the PAA suggest that each of these verbal equating tests is virtually parallel in function to the operational test with which it is paired.

The application of the statistics in Table 4 to equations (1) through (4) and to equations (8) through (11) resulted in the following values:[4]

Tucker Method	*Levine Method*
$\hat{M}_{x_t} = 36.1445$	$\hat{M}_{x_t} = 37.0490$
$\hat{M}_{y_t} = 25.7771$	$\hat{M}_{y_t} = 24.7645$
$\hat{s}_{x_t} = 14.8306$	$n_{xv_a} = 1.6691$
$\hat{s}_{y_t} = 18.2502$	$n_{yv_g} = 2.0578$

From these values the following equations, permitting the conversion of scores from the raw-score scale of the PAA-verbal test (X) to the raw-score scale of the SAT-verbal test (Y), were determined under the Tucker and Levine methods:

$$Y = 1.2306X - 18.7017 \quad \text{(Tucker)}, \tag{16}$$

[3] Recall that 115 verbal "outlier" items were removed from the original group of 155 items administered to both the continental and the Puerto Rican examinees.

[4] Since the PAA and the SAT must be regarded as appropriate only for the cultural group for which each was separately designed, each "combined group" value must be interpreted as an estimate of the performance of the combined group assuming that the test in question was appropriate for all members of the combined group. The Puerto Rican and mainland samples were weighted about equally in making these calculations.

and

$$Y = 1.2329X - 20.9136 \quad \text{(Levine)}. \tag{17}$$

In order to derive the numerical conversion from the PAA-verbal reporting scale to the SAT-verbal reporting scale under the Tucker method, the following numerical values for the slopes and intercepts of equations (6), (13), and (14) were applied to the constants in equation (15):

$$a = 1.2306, \quad b = -18.7017 \quad \text{[from equation (6) by Tucker method]},$$
$$A = 7.1424, \quad B = 264.1965 \quad \text{[from equation (13)]},$$

and

$$A' = 6.3075, \quad B' = 225.4387 \quad \text{[from equation (14)]}.$$

The resulting scale-to-scale conversion for the verbal test, derived by the Tucker method of linear equating, is, therefore,

$$S_c = 1.0868\, S_p - 179.6381. \tag{18}$$

The numerical conversions from the PAA-verbal reporting scale to the SAT-verbal reporting scale under the Levine method were obtained from equations (7), (13), and (14), using the following conversion parameters in a relationship precisely equivalent to that shown in equation (15), except that a' is applied instead of a, and b' instead of b:

$$a' = 1.2329, \quad b' = -20.9136 \quad \text{[from equation (7) by Levine method]},$$
$$A = 7.1424, \quad B = 264.1965 \quad \text{[from equation (13)]},$$

and

$$A' = 6.3075, \quad B' = 225.4387 \quad \text{[from equation (14)]}.$$

The resulting scale-to-scale conversion for the verbal test derived by the Levine method of linear equating now becomes:

$$S_c = 1.0888\, S_p - 194.1262. \tag{19}$$

Comparison of the Tucker and Levine conversions for the verbal tests shows a constant difference of about 13 points throughout the range of scaled scores, with the Levine conversions yielding lower SAT equivalents. This result is predictable from the fact that the PAA group had a lower mean on the equating items than did the SAT group.

Because the data of this study failed to satisfy the assumptions of either the Tucker or the Levine equating methods entirely, the final linear conversion equation for transforming the PAA-verbal scale S_p to the SAT-verbal scale S_c was taken to be the bisector of the two lines given, respectively, in equations (18) and (19). The equation of the bisector is:

$$S_e = 1.0878\, S_p - 186.8779, \qquad (20)$$

from which the following equivalencies were determined:

PAA-V Score	*Equivalent SAT-V Score (Linear)*
800	683
700	575
600	466
500	357
400	248
300	(139)[5]
200	(31)[5]

A more detailed linear conversion table for the verbal tests is provided in Table I of the Appendix (page 35). However, it is clear from the foregoing list of equivalencies (assuming a linear model for equating) that the difference between the two scales is in the vicinity of 140–145 points at a PAA score of 500. The differences are larger, however, at the lower end of the scale and become progressively smaller at the higher score levels.

Directly parallel procedures were followed in deriving the equation for converting scaled scores on the PAA-mathematical sections to scaled scores on the SAT-mathematical sections. In order to calculate the estimated values given in equations (1) through (4) and in equations (8) through (11) for the mathematical tests, the correlations between the operational test and the 25-item equating section,[6] as well as the related means and standard deviations, were prepared for each of the two mathematical samples, one consisting of 3,867 continental students and the other of 1,060 Puerto Rican students. These statistics are given in Table 5 along with the frequency distributions of the mathematical operational and equating tests.

The mathematical equating data in Table 5 reveal even more sharply than do the verbal equating data in Table 4 that the continental sample is the higher scoring of the two. The mean difference in the mathematical "common items" is about 1.5 standard deviations. Also, note that as in the case of the verbal test, the operational PAA-mathematical test was more appropriate in difficulty for the PAA sample (percentage-pass, corrected for guessing = .44) than was the SAT-mathematical test for the SAT sample (percentage-pass, corrected for guessing = .38). The distributions in Table 5, which show moderate positive skewness on the SAT for the continental sample, confirm this observation. The most striking observations to be made from Table 5, however, are the extreme negative skew in the distribution of

[5] Scores lower than 200 on both the SAT and the PAA are reported as 200.

[6] Recall that 75 mathematical "outlier" items were removed from the original group of 100 items administered to both the continental and the Puerto Rican examinees.

TABLE 5 *Frequency Distributions and Summary Statistics for the Operational and Equating Sections of the SAT and PAA*

	Mathematical Tests			
	Continental Sample		Puerto Rican Sample	
Raw (Formula) Score	Operational SAT	Equating Section	Operational PAA	Equating Section
58 – 59	4			
56 – 57	11			
54 – 55	11		9	
52 – 53	21		12	
50 – 51	30		20	
48 – 49	55		25	
46 – 47	49		21	
44 – 45	71		45	
42 – 43	77		30	
40 – 41	99		36	
38 – 39	118		49	
36 – 37	136		24	
34 – 35	163		51	
32 – 33	160		44	
30 – 31	200		49	
28 – 29	197		59	
26 – 27	202		47	
24 – 25	224	882	58	20
22 – 23	197	499	54	21
20 – 21	220	562	61	46
18 – 19	220	453	72	49
16 – 17	161	240	43	45
14 – 15	182	272	57	55
12 – 13	176	208	56	56
10 – 11	166	179	67	68
8 – 9	176	157	29	82
6 – 7	148	118	12	86
4 – 5	120	105	14	131
2 – 3	93	76	6	104
0 – 1	87	63	6	154
−2 – −1	53	41	4	96
−4 – −3	22	8		34
−6 – −5	13	4		13
−8 – −7	4			
−10 – −9	1			
Number of Cases	3,867	3,867	1,060	1,060
Mean	22.6499	17.6025	26.4066	7.1868
SD	13.1246	6.8164	12.7172	7.3849
Correlation: Operational vs. Equating	.8206		.8781	
Number of Items	60	25	55	25

NOTE: The operational PAA-mathematical and SAT-mathematical tests contained 55 and 60 items respectively. The mathematical equating test, administered to both the PAA and SAT groups (in their own language mode), consisted of 25 items.

equating test scores for the continental sample and the positive skew on that test for the Puerto Rican sample—again pointing to the vast difference in difficulty of the "common" mathematics items for these two groups.

As was true of the verbal data in Table 4, the patterns of standard deviations and correlations in Table 5 between the equating test in English and the SAT and between the equating test in Spanish and the PAA suggest that each of these mathematical equating tests is virtually parallel in function to the operational test with which it is paired.

The application of the statistics in Table 5 to equations (1) through (4) and to equations (8) through (11) results in the following values:[7]

Tucker Method	*Levine Method*
$\hat{M}_{x_t} = 35.0494$	$\hat{M}_{x_t} = 36.5940$
$\hat{M}_{y_t} = 15.2236$	$\hat{M}_{y_t} = 13.0176$
$\hat{s}_{x_t} = 14.5954$	$n_{xr_\alpha} = 1.7824$
$\hat{s}_{y_t} = 15.7612$	$n_{yr_\beta} = 2.0494$

These values were then applied to yield the equations for the mathematical tests (corresponding to those for the verbal tests in equations (16) and (17) above), as follows:

$$Y = 1.0799X - 22.6253 \text{ (Tucker)}, \tag{21}$$

and

$$Y = 1.1498X - 29.0573 \text{ (Levine)}, \tag{22}$$

permitting the conversion of scores from the raw-score scale of the PAA-mathematical test to the raw-score scale of the SAT-mathematical test.

In order to derive the Tucker conversion from the PAA-mathematical reporting scale to the SAT-mathematical reporting scale, the following numerical values from the slopes and intercepts of equations (6), (13), and (14) were applied to the constants in equation (15):

$a = 1.0799$, $b = -22.6253$ [from equation (6) by Tucker method],
$A = 8.3361$, $B = 268.2090$ [from equation (13)],

and

$A' = 8.5584$, $B' = 279.2013$ [from equation (14)].

The resulting conversion for the mathematical test under the Tucker method is, therefore,

[7] Since the PAA and the SAT must be regarded as appropriate only for the cultural group for which each was separately designed, each "combined group" value must be interpreted as an estimate of the performance of the combined group assuming that the test in question was appropriate for all members of the combined group. The Puerto Rican and continental samples were weighted about equally in making these calculations.

$$S_c = 1.1087\, S_p - 211.7978. \tag{23}$$

Similarly, the scale-to-scale conversion of the mathematical test obtained under the Levine method by applying the results of equation (22) (namely, $a' = 1.1498$ and $b' = -29.0573$) and the scaled-score parameters A, B, A', and B' in the preceding paragraph to an equation precisely parallel to equation (15) is as follows:

$$S_c = 1.1805\, S_p - 286.0932. \tag{24}$$

Comparison of the Tucker and Levine conversions for the mathematical tests reveals substantial differences, ranging from 60 points at a PAA score of 200 to 17 points at a PAA score of 800, with (as in the verbal conversions) the Levine conversions yielding the lower SAT equivalents. As with the verbal conversions, but to a much more pronounced degree, the difference between the two conversions is predictable from the fact that the PAA group had a lower mean on the equating items than the SAT group did.

As with the verbal conversions, the bisector of the two lines represented by equations (23) and (24) was used as the final conversion line for transforming the PAA-mathematical scale S_p to the SAT-mathematical scale S_c. Its equation is:

$$S_c = 1.1440\, S_p - 248.2851, \tag{25}$$

from which the following equivalencies were determined:

PAA-M Score	*Equivalent SAT-M Score (Linear)*
800	667
700	553
600	438
500	324
400	209
300	(95)[8]
200	(−18)[8]

As do the verbal equivalencies, these (linear) equivalencies for the mathematical tests show striking differences between the PAA and SAT scales. In the vicinity of a PAA score of 500 there is a difference of 175–180 points. However, as with the verbal equivalencies, the differences are larger at the lower end of the scale and become progressively smaller at higher score levels.

A more detailed linear conversion table for the mathematical tests is provided in Table I of the Appendix (page 35).

[8] Scores lower than 200 on both the PAA and SAT are reported as 200.

Curvilinear Equating

The curvilinear (equipercentile) method of equating, outlined above, was also used to determine the equivalent raw scores on PAA and SAT tests. These equivalent scores were obtained by first equating raw scores on test X (administered to the PAA group) to raw scores on the common test V (also given to the same PAA group) by setting equal scores at the same percentile rank on the distributions for tests X and V. Similarly, raw scores at the same percentile ranks on the distributions for tests Y and V (taken by the SAT group) were also equated. Then, for each score on test V, the equivalent scores on tests X and Y were found, plotted, and smoothed to yield a conversion from X to Y.

The equated raw scores on tests X and Y were then converted to their corresponding scaled scores from equations (13) and (14):

$$S_p = AX + B, \tag{13}$$

and

$$S_c = A'Y + B'. \tag{14}$$

These equations, for converting raw scores to scaled scores, are restated in numerical terms as follows:

Verbal

PAA: $A = 7.1424$ $B = 264.1965$
SAT: $A' = 6.3075$ $B' = 225.4387$

Mathematical

PAA: $A = 8.3361$ $B = 268.2090$
SAT: $A' = 8.5584$ $B' = 279.2013$

from which the following equivalencies were determined:

PAA-V Score	*Equivalent SAT-V Score (Curvilinear)*	*PAA-M Score*	*Equivalent SAT-M Score (Curvilinear)*
800	—[9]	800	—[9]
700	628	700	519
600	449	600	386
500	342	500	313
400	260	400	266
300	200	300	215
200		200	

The curvilinear equivalencies tell essentially the same story as do the linear

[9] Score equivalencies above the mid-700s on the PAA are unavailable because of the scarcity of data in the upper region of the distribution of PAA scores.

equivalencies, that there is a wide difference between the PAA and SAT scales. In these equivalencies there is about a 155–160 point difference in verbal scores and about a 185–190 point difference in mathematical scores at a PAA score of 500. Detailed curvilinear conversion tables are provided in Table II of the Appendix (page 36).

The final conversions between the PAA and SAT scales chosen for operational use are the averages of the linear and curvilinear equatings, one for the verbal tests, the other for the mathematical tests. The detailed equivalency tables appear in Table III of the Appendix (page 37). A summary of these equivalencies follows:

PAA-V Score	*Equivalent SAT-V Score (Average)*	*PAA-M Score*	*Equivalent SAT-M Score (Average)*
800	767[10]	800	—[11]
700	602	700	536
600	458	600	412
500	350	500	319
400	254	400	238

Graphs of the linear and curvilinear equatings, as well as the final conversions (which are the averages of the two), appear in Figures 3 and 4.

The essentials of the relationships between the score scales for the PAA and the SAT that are observed in these final equivalency tables have already been described in connection with the linear and equipercentile results presented earlier in this report. The tables indicate that a PAA midscale value (500) is equivalent to an SAT-verbal score substantially below midvalue (350), and an even lower (319) SAT-mathematical score.

Some attention should be given to the meaning of the differences in these scales. The fact that a 500 score on the PAA corresponds to a lower-than-500 score on the SAT simply says that if one can assume that the SAT and PAA values have been maintained precisely since the time of their inception, it can be concluded that the original scaling group for the SAT was generally more able in the abilities measured by these aptitude tests than the original scaling group for the PAA. It does not by itself imply that the SAT candidate group today is necessarily more able than the PAA group, although this is in fact the case; the 1971–72 PAA candidate group earned mean scores on their own scale of 478 on the verbal test and 484 on the mathematical test, which convert, respectively, to about 328 and 304 on the SAT scale. The 1971–72 SAT candidate group earned considerably higher mean scores —450 on the verbal test and 482 on the mathematical test. Nor does it necessarily suggest any generalization regarding the larger populations from which these two examinee groups were self-selected—for example, that the twelfth-grade students

[10] Extrapolated value.

[11] Further extrapolation was not possible in this region.

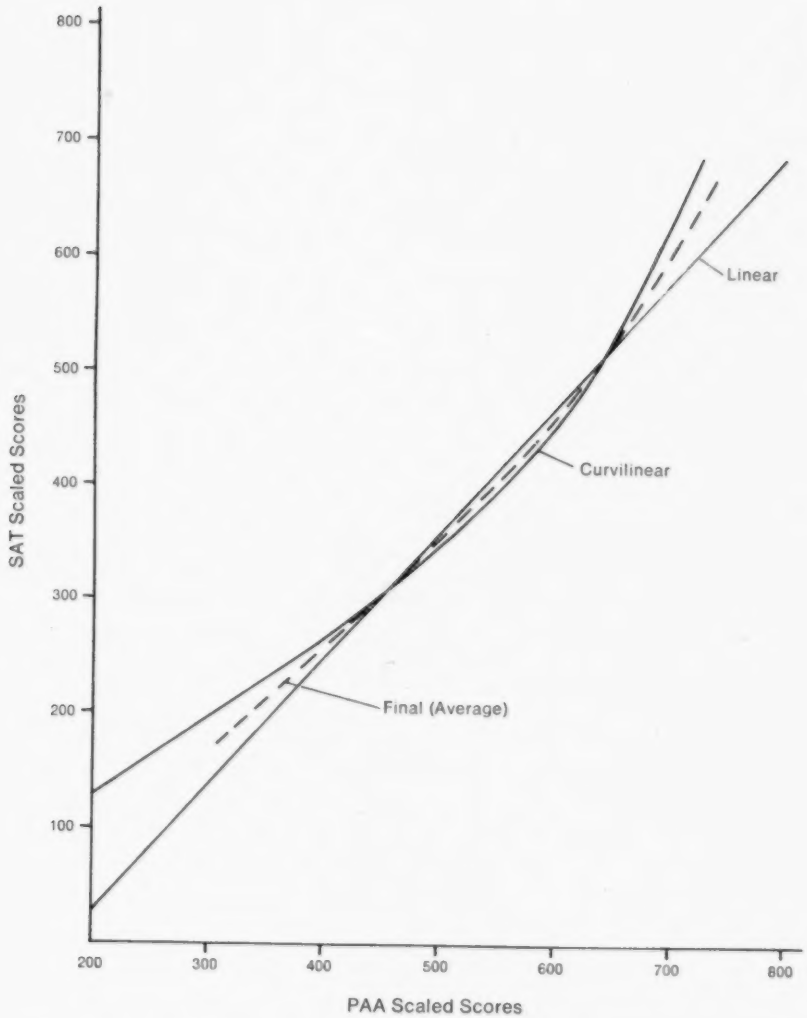

FIGURE 3. *Linear, Curvilinear, ana Final (Average) Conversions for the Verbal Tests*

on the mainland are higher scoring than the twelfth-grade students in Puerto Rico. We know, for example, that the SAT examinee group represents about one-third of the twelfth-grade population on the mainland and is therefore a more selective group than its PAA counterpart, which represents a larger proportion, about two-thirds, of the twelfth-grade population in Puerto Rico. On the other hand, this is not to say that differences between the two twelfth-grade populations do not also exist. There is some evidence, however crude, that marked differences do exist. But this evidence is outside the scope of the present study.

In view of these and other possible misinterpretations of the data of this study, it will be useful to restate the limited purpose for which the present investigation was undertaken: to derive a set of conversions between two similar-appearing scales of measurement, one for tests of one language and culture, the other for

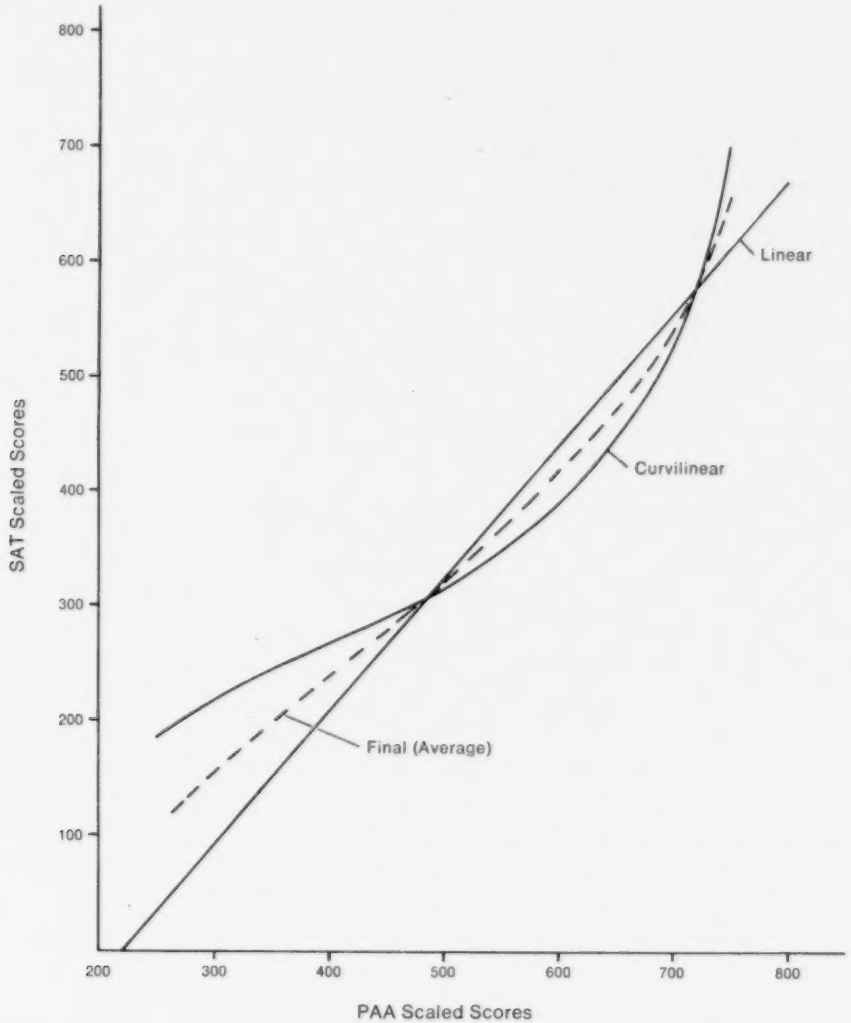

FIGURE 4. *Linear, Curvilinear, and Final (Average) Conversions for the Mathematical Tests*

tests of a different language and culture. Clearly, the accuracy of these conversions is limited by the appropriateness of the method used to derive them and the data assembled during the course of the study. It is hoped that these conversions will be useful in a variety of contexts but (as suggested by the examples cited here) in order to be useful they will need in each instance to be supported by additional data peculiar to the context.

Summary and Discussion

The purpose of this study was to establish score equivalencies between the College Board Scholastic Aptitude Test (SAT) and its Spanish-language equivalent, the

College Board Prueba de Aptitud Académica (PAA). The method of the study involved two phases: the selection of test items equally appropriate and useful for Spanish- and English-speaking students for use in the equating of the two tests; and the equating analysis itself. The method of the first phase was to choose two sets of items, one originally appearing in Spanish, the other originally appearing in English; to translate each set into the other language; and to administer both sets in the appropriate language mode for pretest purposes to both types of students. These administrations were conducted in the fall of 1970 with samples of candidates taking the PAA or the SAT at regularly scheduled administrations. They provided data regarding the difficulty and discrimination power of each item for each of the two groups and, what was of special interest, an index of the appropriateness of each item for each group.

On the basis of the analyses of these data, two sets of items, one verbal and the other mathematical, were chosen and assembled as "common items" to be used for equating. In the second phase of the study these "common items," appearing both in Spanish and in English, were administered in the appropriate language along with the operational form of the PAA in November 1971 and with the operational form of the SAT in January 1972. The data resulting from the administrations of these "common items" were used to calibrate for differences in the abilities of the two groups of candidates, and permitted both linear and curvilinear (equipercentile) equating of the two tests. Conversion tables relating the PAA-verbal scores to the SAT-verbal scores and the PAA-mathematical scores to the SAT-mathematical scores are given in the Appendix (page 37). These conversions represent an average of the linear and equipercentile results. Because of the scarcity of data at the upper end of the distribution of PAA scores, score equivalencies are permissible, strictly speaking, only as high as the mid-700s. Score equivalencies beyond the mid-700s were obtained by extrapolation.

The procedure followed in conducting this study requires special discussion, perhaps all the more because it is, at least superficially, a simple one both in its conception and in its execution. On the other hand, from a psychological viewpoint the task of making cross-cultural comparisons of the kind made here is highly complex. In the extreme the task is inescapably an impossible one, and although the present study may represent a reasonably successful attempt, it should be remembered that the cultural differences confronted by the present study were minimal and relatively easily bridged. If, for example, the two cultures under consideration were very different, then there would be little or no common basis for comparison.

Given, then, that the cultures out of which the tests in the present study were developed are to some extent similar, and that there is indeed a basis for comparison, the approach and method offered in this study do appear to have some likelihood of success. Indeed, the method itself is useful not only in providing a type of metric for utilizing the common basis for comparison, but also in providing a basis for evaluating the degree to which there is a common basis for comparison. For example, it allows a comparison of the two cultures only on a common ground, which is to say only on those items that are relatively close to the major axis of

the ellipse. This being the case, those characteristics of the two cultures that make them uniquely different are in essence removed from consideration in making the comparisons. Thus, while we are afforded an opportunity to compare the two cultures on a common basis—i.e., on the items that are "equally appropriate"—at the same time we are afforded an opportunity to examine the differences in the two cultures in the terms provided by the divergent or "unequally appropriate" items. It is noteworthy that what emerges out of this study—and other studies that have also made use of the delta-plot technique, e.g., Angoff and Ford (1971)—is that the method described here also yields a general measure of cultural similarity, expressed in the size of the correlation represented by the delta plots. The correlation (or statistics derived from the correlation, such as the standard deviation of the *D*-values) summarizes the degree to which members of the two cultures perceive the item stimuli similarly. Additional studies of the similarity of any two cultures would have to be based on other stimuli examined in a wide variety of different social contexts.

It should also be made clear that the method has its limitations, as do the results of this study which has followed the method. For example, the present study has leaned on the usefulness of translations from each of the two languages to the other, and the assumption has been made that biases in translation, if they exist, tend to balance out. This assumption may not be a tenable one, however. Quite possibly translation may be easier and freer of bias when going from language A to language B than in the reverse direction; and if items do become somewhat more difficult in an absolute sense as a result of translation, this effect would be more keenly felt by speakers of language A than of language B. The result of this effect is that the central tendency of the elliptical plot of common items would experience a net bias. Also, implicit in the method of this study is the assumption that language mirrors all the significant cultural effects. This may not be so, and it is possible that the translatability of words and concepts across two languages does not accurately reflect the degree of similarity in the cultures represented by those two languages. If, for example, there are greater differences in the languages than in the cultures—as may perhaps be the case between the German and Hungarian languages as compared with the German and Hungarian cultures—then again the method is subject to some bias.

Aside from matters of methodology and possible sources of bias, a point that has been made earlier in this report deserves repeating: The comparison in this study was made between Puerto Rican and continental U.S. students; the resulting conversions between the PAA and the SAT apply only between these two groups of students. Whether the same conversions would also have been found had the study been conducted between the PAA and the SAT as taken by other Spanish speakers and other English speakers is an open question. Indeed, it is an open question whether the conversion obtained here also applies to variously defined subgroups of the Puerto Rican and continental populations—liberal arts women, engineering men, urban blacks, etc.

It is also to be hoped that the conversions between the two types of tests will not be used without a clear recognition of the realities: A Puerto Rican student

with a PAA-verbal score of 680 has a score "equivalent" to an SAT-verbal score of 568. This is not to say that that student could actually earn an SAT-verbal score of 568 were he to take the SAT. He might do better, or he might do worse, depending, obviously, on his facility in English. The conversions do offer a way of evaluating his general aptitude for verbal and mathematical materials in terms familiar to users of SAT scores; and, depending on how well he can be expected to learn the English language, his likelihood of success in competition with native English speakers in the continental United States can be estimated. Continuing study of the comparative validity of the PAA and the SAT for predicting the performance of Puerto Rican students in mainland colleges is indispensable to the judicious use of these conversions.

References

Angoff, W. H. "Test Reliability and Effective Test Length." *Psychometrika*, March 1953, *18*, No. 1, pp. 1–14.

Angoff, W. H. "Can Useful General-Purpose Equivalency Tables Be Prepared for Different College Admissions Tests?" In A. Anastasi (Ed.), *Testing Problems in Perspective*. Washington, D.C.: American Council on Education, 1966. Pp. 251–264.

Angoff, W. H. "Scales, Norms, and Equivalent Scores." In R. L. Thorndike (Ed.), *Educational Measurement*. (2nd ed.) Washington, D.C.: American Council on Education, 1971. Pp. 508–600.

Angoff, W. H., and Ford, S. F. "Item-Race Interaction on a Test of Scholastic Aptitude." College Entrance Examination Board Research and Development Report 71–72, No. 3. Princeton, N.J.: Educational Testing Service, 1971.

Boldt, R. F. "Concurrent Validity of the PAA and SAT for Bilingual Dade County High School Volunteers." College Entrance Examination Board Research and Development Report 68–69, No. 3. Princeton, N.J.: Educational Testing Service, 1969.

Levine, R. S. "Equating the Score Scales of Alternate Forms Administered to Samples of Different Ability." Research Bulletin 55–23. Princeton, N.J.: Educational Testing Service, 1955.

Thurstone, L. L. "The Calibration of Test Items." *American Psychologist*, 1947, *2*, 103–104.

Appendix

TABLE I *Linear Conversions of PAA Scaled Scores to SAT Scaled Scores*

Verbal				Mathematical			
PAA-V Score	Equivalent SAT-V Score	PAA-V Score	Equivalent SAT-V Score	PAA-M Score	Equivalent SAT-M Score	PAA-M Score	Equivalent SAT-M Score
800	683	490	346	800	667	490	312
		480	335			480	301
790	672	470	324	790	655	470	289
780	662	460	314	780	644	460	278
770	651	450	303	770	633	450	267
760	640	440	292	760	621	440	255
750	629	430	281	750	610	430	244
740	618	420	270	740	598	420	232
730	607	410	259	730	587	410	221
720	596	400	248	720	575	400	209
710	585			710	564		
700	575	390	237	700	553	390	(198)
		380	226			380	(186)
690	564	370	216	690	541	370	(175)
680	553	360	205	680	530	360	(164)
670	542	350	(194)	670	518	350	(152)
660	531	340	(183)	660	507	340	(141)
650	520	330	(172)	650	495	330	(129)
640	509	320	(161)	640	484	320	(118)
630	498	310	(150)	630	472	310	(106)
620	488	300	(139)	620	461	300	(95)
610	477			610	450		
600	466	290	(129)	600	438	290	(83)
		280	(118)			280	(72)
590	455	270	(107)	590	427	270	(61)
580	444	260	(96)	580	415	260	(49)
570	433	250	(85)	570	404	250	(38)
560	422	240	(74)	560	392	240	(26)
550	411	230	(63)	550	381	230	(15)
540	401	220	(52)	540	369	220	(3)
530	390	210	(42)	530	358	210	
520	379	200	(31)	520	347	200	
510	368			510	335		
500	357			500	324		

NOTE: The lowest score reported on both the PAA and the SAT is the score of 200.

TABLE II *Curvilinear Conversions of PAA Scaled Scores to SAT Scaled Scores*

Verbal				Mathematical			
PAA-V Score	Equivalent SAT-V Score	PAA-V Score	Equivalent SAT-V Score	PAA-M Score	Equivalent SAT-M Score	PAA-M Score	Equivalent SAT-M Score
800		490	334	800		490	305
		480	324			480	301
790		470	317	790		470	296
780		460	307	780		460	292
770		450	299	770		450	288
760		440	290	760		440	283
750		430	282	750		430	279
740		420	275	740	660	420	272
730	686	410	268	730	598	410	268
720	666	400	260	720	570	400	266
710	648			710	545		
700	628	390	252	700	519	390	258
		380	246			380	255
690	606	370	240	690	502	370	250
680	583	360	234	680	482	360	245
670	565	350	227	670	467	350	241
660	541	340	220	660	455	340	236
650	525	330	213	650	439	330	230
640	503	320	207	640	431	320	224
630	487	310	201	630	418	310	219
620	475	300		620	409	300	215
610	459			610	396		
600	449	290		600	386	290	207
		280				280	202
590	438	270		590	378	270	(198)
580	424	260		580	367	260	
570	413	250		570	361	250	
560	402	240		560	353	240	
550	393	230		550	344	230	
540	383	220		540	339	220	
530	372	210		530	331	210	
520	363	200		520	320	200	
510	353			510	320		
500	342			500	313		

NOTE: The lowest score reported on both the PAA and the SAT is the score of 200.

TABLE III *Final Conversions Between PAA Scaled Scores and SAT Scaled Scores*

Verbal				Mathematical			
PAA-V Score	Equivalent SAT-V Score	PAA-V Score	Equivalent SAT-V Score	PAA-M Score	Equivalent SAT-M Score	PAA-M Score	Equivalent SAT-M Score
800	767[a]			800	—[b]		
		490	340			490	309
790	750[a]	480	330	790	—[b]	480	301
780	733[a]	470	321	780	—[b]	470	293
770	715[a]	460	311	770	—[b]	460	285
760	698[a]	450	301	760	—[b]	450	278
750	680[a]	440	291	750	650[a]	440	269
740	663[a]	430	282	740	629	430	262
730	647	420	273	730	593	420	252
720	631	410	264	720	573	410	245
710	617	400	254	710	555	400	238
700	602			700	536		
690	585	390	245	690	522	390	228
680	568	380	236	680	506	380	221
670	554	370	228	670	493	370	213
660	536	360	220	660	481	360	205
650	523	350	211	650	467	350	(197)
640	506	340	202	640	458	340	(189)
630	493	330	(193)	630	445	330	(180)
620	482	320	(184)	620	435	320	(171)
610	468	310	(176)	610	423	310	(163)
600	458	300		600	412	300	(155)
590	447	290		590	403	290	(145)
580	434	280		580	391	280	(137)
570	423	270		570	383	270	(131)
560	412	260		560	373	260	
550	402	250		550	363	250	
540	392	240		540	354	240	
530	381	230		530	345	230	
520	371	220		520	337	220	
510	361	210		510	328	210	
500	350	200		500	319	200	

NOTE: The lowest score reported on both the PAA and the SAT is the score of 200. For operational use these conversions should be rounded to the nearest multiple of 10.

[a] Extrapolated values.

[b] Further extrapolation was not possible in this region.